

# L0 regularization for the estimation of piecewise constant hazard rates in survival analysis

Olivier Bouaziz<sup>1</sup> and Grégory Nuel<sup>2</sup>

<sup>1</sup>Laboratory MAP5, University Paris Descartes and CNRS, Sorbonne Paris Cité,  
Paris, France

<sup>2</sup>LSTA, CNRS 7599, 4 place Jussieu, F-75005 Paris, France

## Abstract

In a survival analysis context we suggest a new method to estimate the piecewise constant hazard rate model. The method provides an automatic procedure to find the number and location of cut points and to estimate the hazard on each cut interval. Estimation is performed through a penalized likelihood using an adaptive ridge procedure. A bootstrap procedure is proposed in order to derive valid statistical inference taking both into account the variability of the estimate and the variability in the choice of the cut points. The new method is applied both to simulated data and to the Mayo Clinic trial on primary biliary cirrhosis. The algorithm implementation is seen to work well and to be of practical relevance.

**Keywords:** Adaptive Ridge procedure; Hazard rate estimation; Penalized likelihood; Piecewise constant hazard; Survival analysis.

## 1 Introduction

In survival analysis, when interest lies on the estimation of the hazard rate, an attractive and popular model is the piecewise constant hazard model. This model is easy to interpret as the hazard rate is supposed to be constant on some pre-defined time intervals and plotting the hazard rate gives a quick sense of the evolution of the event of interest through time. Many epidemiological studies use this model to represent the hazard rate function either because it provides an interesting way to fit the hazard function or because the data are not available on the individual level. See for instance Table 1 of Antoniou et al. (2004) where the authors displayed the incidence of breast cancer on ten-year intervals for different subpopulations.

While this model can be used in a nonparametric setting, it is often used in combination with covariates effects. This is the case for instance for the popular Poisson regression model (see Clayton et al. (1993) or Aalen et al. (2008)) which assumes a proportional effect on the covariates and a piecewise constant hazard model for the baseline

hazard. This model is widely used in practice typically when dealing with register data. On one hand it allows to perform survival analysis with large computational savings (and save considerable data storage requirements) and, on the other hand, it allows to easily estimate the baseline hazard rate as a piecewise constant function and to give a very easy interpretation of the baseline hazard rate. Among many practical examples, we refer the reader to Kessing et al. (2010), Jensen et al. (2013) or Hviid & Svanström (2009). In practice, as noticed by Grøn et al. (2016) for Poisson regression, *“the choice of time intervals should generally be guided by subject matter aspects, but the numbers of events and numbers at risk within intervals may also be considered when specifying the number and lengths of the intervals. A study of a rare event and/or a small exposure group may require longer intervals.”* While this might be true, it is clear that in some situations there might be no a priori knowledge for the choice of the time intervals and then they are usually arbitrarily chosen. This is the case for example in Antoniou et al. (2004) where the time intervals in Table 1 were arbitrarily chosen as ten years length.

When modeling covariates effect through a proportional hazard model, Cox (1972) proposed an estimator that allows the baseline to stay unspecified. In this model, the baseline is taken as a function that only puts mass on the observed events and the likelihood simplifies into the Cox partial likelihood where the regression effect can be estimated separately from the baseline. While this is a very interesting aspect of the Cox model, this nice separation between baseline estimation and regression effect estimation does not hold anymore in many extensions of this model. For instance, in frailty models (see among many other authors Clayton (1978), Hougaard (1995), Therneau & Grambsch (2000) and Ripatti & Palmgren (2002)) keeping a non-parametric baseline makes the estimation method much more complicated since baseline and regression parameters must be estimated simultaneously. As a consequence, the literature on estimation procedures in the frailty context is vast. As a matter of fact the estimation procedures in Klein et al. (1992), Andersen et al. (1997) and Ripatti & Palmgren (2002) all lead to similar but still different estimates. Importantly, in Andersen et al. (1997) it is said that Poisson regression and Cox models give results that tend to be very similar, with or without frailties.

In the joint modeling framework one wants to model the association between a longitudinal variable and a time to event response through a random effect (see Tsiatis & Davidian (2004), Rizopoulos (2012)). Only parametric baseline functions are implemented in the widely used `jm` R package (see Rizopoulos et al. (2010)). As a matter of fact, the author in Rizopoulos (2012) recommends either to use the piecewise constant baseline hazard or a spline basis baseline hazard which he says *“often work quite satisfactorily in practice”* (see page 53 of the book). The R `frailtypack` package (see Rondeau et al. (2012)) deals with more survival analysis situations involving a random effect such as nested frailty models (see Rondeau et al. (2006)) or joint inference of recurrent and terminal events (see Rondeau et al. (2007)). In this package, the possible baseline hazard functions are the piecewise constant hazard, Weibull hazard and spline functions. In the last case, the authors introduce a penalized likelihood estimation method that allows to obtain smooth estimates of the baseline hazard function. However the use of

spline baseline functions requires to specify in advance the number of knots used in the estimation and therefore can be seen as a smoothed version of the piecewise constant hazard functions where one must choose in advance the number of cuts.

Other contexts where the partial likelihood approach does not work anymore include the cure models framework (see for instance Farewell (1982) and Sy & Taylor (2000)) and the analysis of interval-censoring data (see Sun (2007) for instance). In the latter case, the nonparametric maximum likelihood estimator for the cumulative hazard or the survival function is known to be slow with a convergence rate of order  $n^{-1/3}$  and the limiting distribution is not Gaussian (see Groeneboom & Wellner (1992) for current status data and Groeneboom (1996) for case II intervals censored data). This problem pertains in the regression framework (see sections 5.2.3 and 6.2.2 in Sun (2007) for instance). On the other hand, using parametric baseline functions such as the piecewise hazard functions allows to obtain classical parametric rate of convergence and makes the estimation procedure much more stable.

In this article, we only consider the nonparametric setting of estimating the baseline hazard function in a piecewise constant hazard model in the situation of right-censored data. We propose a new method to automatically find the appropriate number and location of the cuts used in this model. Our algorithm is based on the recent work from Frommlet & Nuel (2016) where starting from a large set of possible cut points an L0 penalty on the likelihood of the model forces many successive cuts to be equal providing a parsimonious estimate of the hazard function. The procedure is data-driven and inference taking into account both the variability from the estimates and the cut points positions can be derived.

In Section 2 the piecewise constant hazard model is recalled and the adaptive ridge estimator is applied to this model. Section 3 proposes two different procedures to choose the penalty term involved in the estimation procedure. Section 4 proposes a bootstrap based method to obtain valid inference for survival distribution quantities such as the survival function. A simulation study is conducted in Section 5, where the efficiency of the estimation method is evaluated and the two different procedures to choose the penalty term are compared. The method is applied to the Mayo Clinic trial on primary biliary cirrhosis in Section 6 and a small discussion concludes the paper in Section 7.

## 2 Model and estimation procedure

### 2.1 The piecewise constant hazard rate model

Let  $T^*$  represent the survival time of interest. In practice  $T^*$  might be censored by a random variable  $C$  so that we observe  $(T = T^* \wedge C, \Delta = I(T^* \leq C))$ . Let  $\tau$  be the endpoint of the study, the data consist of  $n$  independent replications  $(T_i, \Delta_i)_{i=1, \dots, n}$ . We aim at estimating the hazard function defined for  $t \in [0, \tau]$  by:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}[t \leq T^* < t + \Delta t | T^* \geq t]}{\Delta t}.$$

In the following, this hazard function is assumed to be piecewise constant on  $L$  cuts represented by  $c_0, c_1, \dots, c_L$ , with the convention that  $c_0 = 0$  and  $c_L = +\infty$ . Let  $I_l(t) = I(c_{l-1} < t \leq c_l)$ . We suppose that

$$\lambda(t) = \sum_{l=1}^L I_l(t) \alpha_l,$$

for  $\alpha_l \geq 0, l = 1, \dots, L$ . Note that the exponential baseline hazard is obtained from  $L = 1$  in the piecewise constant hazard family.

Let  $\Lambda(t) = \int_0^t \lambda(s) ds$  represents the cumulative hazard function. We denote by  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_L)$  the model parameter we aim to estimate.

In order to make inference on the model parameter we will assume that the endpoint  $\tau$  is defined such that, for all  $t$  in  $[0, \tau]$ ,  $\mathbb{P}[T > t] > 0$ . This assumption is common in survival analysis settings to prevent usual estimation problems that occur in the right tail of the distribution of  $T$ . We will also assume independent right censoring and non-informative censoring (see Andersen et al. (1993) for instance for a complete review of these assumptions). Estimation is then carried out using classical likelihood arguments.

Let  $L_n(\boldsymbol{\alpha}) = \log \prod_{i=1}^n \mathbb{P}[T_i, \Delta_i; \boldsymbol{\alpha}]$  represents the log-likelihood of the model. We have:

$$L_n(\boldsymbol{\alpha}) = \sum_{i=1}^n \left\{ \log(\lambda(T_i)) \Delta_i - \int_0^{T_i} \lambda(t) dt \right\},$$

where the equality holds true up to a constant that does not depend on the model parameter  $\boldsymbol{\alpha}$ . For computational purpose, it is interesting to note that the log-likelihood can be written in a Poisson regression form. Introduce  $R_{i,l} = I(T_i \geq c_{l-1})(c_l \wedge T_i - c_{l-1})$ , the total time individual  $i$  is at risk in the  $l$ th interval  $(c_{l-1}, c_l]$ ,  $O_{i,l} = I_l(T_i) \Delta_i$ , the number of events for individual  $i$  in the  $l$ th subinterval. Also  $R_l = \sum_{i=1}^n R_{i,l}$  and  $O_l = \sum_{i=1}^n O_{i,l}$  are sufficient statistics and estimation can be carried out using only these two statistics. The log-likelihood can then be written again as (see Aalen et al. (2008) p.223-225 for more details):

$$L_n(\boldsymbol{\alpha}) = \sum_{l=1}^L \{O_l(\log(\alpha_l) - \alpha_l R_l)\}. \quad (1)$$

Since  $L_n$  is concave, the maximum likelihood estimator has an explicit solution, obtained from derivation of the log-likelihood: for  $l = 1, \dots, L$ ,

$$\hat{\alpha}_l = \frac{O_l}{R_l}. \quad (2)$$

## 2.2 The adaptive ridge regression

For computational purpose, introduce  $a_l$  such that  $\alpha_l = \exp(a_l)$  and  $\mathbf{a} = (a_1, \dots, a_L)^T$  the vector parameter we aim to estimate. Using the L0 penalty from Frommlet & Nuel

(2016), we propose the following penalized likelihood:

$$L_n^{\text{pen}}(\mathbf{a}, \mathbf{w}) = \sum_{l=1}^L \{O_l a_l - \exp(a_l) R_l\} - \frac{\text{pen}}{2} \sum_{l=1}^{L-1} w_l (a_{l+1} - a_l)^2, \quad (3)$$

where  $\mathbf{w} = (w_1, \dots, w_{L-1})$  are non-negative weights that will be iteratively updated in order for the weighted ridge penalty term to approximate the L0 penalty.

The score vector is denoted  $U(\mathbf{a}, \mathbf{w}) = \partial L_n^{\text{pen}}(\mathbf{a}, \mathbf{w}) / \partial \mathbf{a}$  and its  $l$ th component,  $l \in \{1, \dots, L\}$ , is equal to:

$$O_l - R_l \exp(a_l) + (w_{l-1} a_{l-1} - (w_{l-1} + w_l) a_l + w_l a_{l+1}) \text{pen},$$

with the convention  $w_0 = w_L = a_0 = a_{L+1} = 0$ . Now introduce  $I(\mathbf{a}, \mathbf{w}) = -\partial^2 U(\mathbf{a}, \mathbf{w}) / \partial \mathbf{a}^T$ , the opposite of the Hessian matrix.  $I(\mathbf{a}, \mathbf{w})$  is a  $L \times L$  non-negative definite band matrix whose bandwidth equals 1. Its diagonal elements are equal to

$$I(\mathbf{a}, \mathbf{w})_{l,l} = R_l \exp(a_l) + (w_{l-1} + w_l) \text{pen},$$

other elements next to the diagonal are defined for  $l = 1, \dots, L-1$  by

$$I(\mathbf{a}, \mathbf{w})_{l,l+1} = I(\mathbf{a}, \mathbf{w})_{l+1,l} = -w_l \text{pen},$$

and all other elements are equal to zero, that is for  $l, l'$  such that  $|l-l'| \geq 2$ ,  $I(\mathbf{a}, \mathbf{w})_{l,l'} = 0$ .

The vector parameter  $\mathbf{a}$  is updated using the Newton-Raphson algorithm. For a given sequence of weights  $\mathbf{w}^{(m-1)}$  obtained at the  $(m-1)$ th step, the  $m$ th Newton Raphson iteration step is obtained from the equation

$$\mathbf{a}^{(m)} = \mathbf{a}^{(m-1)} + I(\mathbf{a}^{(m-1)}, \mathbf{w}^{(m-1)})^{-1} U(\mathbf{a}^{(m-1)}, \mathbf{w}^{(m-1)}).$$

The inversion of the band matrix is performed through a fast (linear complexity) C++ implementation of the well-known LDL algorithm (variant of the LU decomposition for symmetric matrices). Initialization of the Newton Raphson algorithm can be obtained from the classical unpenalised estimator of the piecewise constant hazard model, that is  $\mathbf{a}^{(0)} = \arg \max_{\mathbf{a}} L_n(\mathbf{a})$ . See Aalen et al. (2008) for details about this estimator.

On the other hand, following the recommendation from Frommlet & Nuel (2016), the weights can be updated from the equation

$$w_l^{(m)} = \left( (a_{l+1}^{(m)} - a_l^{(m)})^2 + \delta^2 \right)^{-1},$$

for  $l = 1, \dots, L-1$  with  $\delta = 10^{-5}$ . Briefly, this form of the weights is motivated by the fact that  $w_l (a_{l+1} - a_l)^2$  is close to 0 when  $|a_{l+1} - a_l| < \delta$  and close to 1 when  $|a_{l+1} - a_l| > \delta$ . Hence the penalty term tends to approximate the L0 norm. The weights are initialized by  $w_l^{(0)} = 1$ , which gives the standard ridge estimate of  $\mathbf{a}$ .

### 3 Choice of the penalty term

In this section we propose two different ways to choose the correct penalty term. The first one is based on a standard cross-validation criterion while the second one is based on a BIC criterion.

In order to choose the right penalty term, one must first define a large grid of penalty values such that the criterion (cross-validation or BIC) will be evaluated at each of these penalty terms. For that purpose, the algorithm can benefit from a warm start of the penalty weights. Indeed, instead of initializing the weights to 1 for each penalty value, one can take the final weights of the previous (smaller) penalty as a starting point for the next (larger) penalty. In this way, full regularization path similar to those of the LASSO can be generated very efficiently. Note, however, that this warm-starting is not necessary since it is always possible to initialize the algorithm with neutral weights of value 1. A preliminary set of cut positions must also be chosen. For simplicity we recommend to take a large set of equally spaced points including the range of the observed time point values. See Sections 5 and 6 to see how this works in practice.

#### 3.1 A cross-validation criterion

Split the data in  $k$  pieces and define  $\hat{\mathbf{a}}_{\text{pen}}^{-I}$  as the maximizer of the penalized likelihood in Equation (3) when part  $I$  is left out from the data.

Then the  $k$ -fold cross validated log-likelihood is defined by:

$$cv(\text{pen}) = \sum_I L_I(\hat{\mathbf{a}}_{\text{pen}}^{-I}),$$

where  $L_I$  represents the unpenalized log-likelihood as in Equation (1) but computed only in part  $I$  of the data. Maximizing this quantity with respect to pen gives the optimal penalty term.

Note that unlike the Cox regression framework where the baseline is left unspecified, this cross-validated criterion is well defined since in our case the hazard rate is constructed as a continuous function of time. Also, the relation

$$\sum_I L_I(\hat{\mathbf{a}}_{\text{pen}}^{-I}) = \sum_I \left\{ L_n(\hat{\mathbf{a}}_{\text{pen}}^{-I}) - L_{-I}(\hat{\mathbf{a}}_{\text{pen}}^{-I}) \right\}$$

holds such that our criterion is completely equivalent to the cross-validated criterion developed by van Houwelingen et al. (2006) and Simon et al. (2011) in the standard Cox regression framework.

In order to improve efficiency and time speed in the computation programs, the 10-fold cross validation is recommended in practice.

#### 3.2 A BIC criterion

The following criterion can be used as an alternative to the choice of the penalty term. It is defined as a balance between good fit of the data and low complexity of the model

parameters. It is fast to compute and has the following expression:

$$BIC(\text{pen}) = -2L_n(\hat{\mathbf{a}}_{\text{pen}}) + d \log(n).$$

The parameter estimator  $\hat{\mathbf{a}}_{\text{pen}}$  is defined as the maximizer of the penalized likelihood in Equation (3) while  $d$  represents the model complexity. It is equal to the number of distinct consecutive values of the  $a_l$ s in  $\hat{\mathbf{a}}_{\text{pen}}$ :

$$d = \sum_{l=0}^{L-1} I(\hat{a}_{l+1,\text{pen}} - \hat{a}_{l,\text{pen}} \neq 0),$$

with the convention  $a_0 = 0$ .

The performance in the choice of the penalty term by both criteria is investigated in the simulation study in Section 5.

## 4 Statistical inference for the time to event distribution

In practice it is of interest to derive confidence intervals for marginal quantities directly related to the time to event variable such as the cumulative hazard function or the survival function. Asymptotic properties of the piecewise-constant hazard model for a given set of cut points is straightforward and has been already derived (see for instance Aalen et al. (2008)). However, the adaptive ridge estimator involves data driven choice of the cut points and using standard results to derive pointwise confidence intervals for the survival function for instance would lead to an overfitting of this function. This is of major concern and one should interpret such confidence intervals with caution.

One way to take into account the uncertainty in the choice of the cut points is to use a resampling technique where for each sample a different penalty term is chosen from the cross-validated or BIC criterion. This will provide a new hazard estimate with a different set of cut points for each sample. Taking the adequate quantile at each time point allows us to obtain pointwise confidence intervals of the correct order for the quantity of interest.

Interestingly, this resampling technique also allows us to compute an alternative pointwise estimate of the survival function (or of any marginal distribution quantity) by taking the pointwise medians of each bootstrap sample. This provides a very smooth estimate function and, in that sense, this kind of estimate can be seen as a smooth non-parametric estimate of the survival function.

This bootstrap procedure is illustrated in Sections 5 and 6 to derive confidence intervals and smooth estimates for the survival function.

## 5 Simulation study

### 5.1 Simulations under a piecewise constant hazard model

We illustrate the proposed method to estimate the following hazard function:

$$\lambda(t) = \begin{cases} 0 & \text{for } t \in [0, 20], \\ 0.5 \cdot 10^{-2} & \text{for } t \in (20, 40], \\ 1 \cdot 10^{-2} & \text{for } t \in (40, 50], \\ 2 \cdot 10^{-2} & \text{for } t \in (50, 70], \\ 4 \cdot 10^{-2} & \text{for } t > 70. \end{cases}$$

The censoring distribution is simulated as a uniform distribution over the time interval  $[70, 90]$  which gives on average 62% of observed failures. On average, 9.5% of the observations fall into the interval  $(20, 40]$ , 8.5% of the observations fall into the interval  $(40, 50]$ , 27% of the observations fall into the interval  $(50, 70]$  and 55% of the observations fall into the interval  $(70, +\infty)$ .

We start with a single sample of size 100 generated from this model. Using the true cuts, the classical unpenalized hazard estimator derived from Equation (2) is computed on Figure 1. The estimation is quite accurate on each cut interval. Figure 2 presents the two extreme situations where the penalized hazard estimate is computed using a very small penalty term on the left panel and using a very large penalty term on the right panel. We see that in the left panel the hazard function is overfitted while the right panel corresponds to the exponential model. A good choice of the penalty term should provide a good compromise between these two situations. The set of all possible cuts was chosen as all the integer values ranging from 1 to 100 and the set of penalty terms was taken, on the log scale, as the set of 100 equally spaced values ranging from  $\log(0.1)$  to  $\log(1000)$ . On this sample the BIC and cross-validation criteria respectively chose the penalty values equal to 0.95 and 1.15 which both gave the same estimate. Figures 3 shows the regularization path for the penalty term and the penalized estimated hazard obtained from the penalty equal to 0.95. We see that both criteria find only three cuts in the estimation of the hazard function, and the cut interval  $(40, 50]$  is not found by the method on this example. As an indicator of the estimation accuracy, the total variation distance between the true hazard and the penalized hazard estimate is computed on the time interval  $[0, 80]$ . On our data example, the total variation is approximately equal to 0.29. Finally, confidence intervals are derived for the survival function using the resampling technique presented in Section 4. The curves are plotted in Figure 4 from 100 bootstrap samples. Our method shows very little difference from the classical Kaplan-Meier estimate and its pointwise confidence interval. Interestingly, our survival estimator and its pointwise confidence intervals have a smooth shape in contrast with the stepwise shape of the Kaplan-Meier estimator.

In order to assess the good performance of our penalized estimator, we also conducted Monte-Carlo simulations from the model scenario presented in this section with 600 sample replications. We considered samples of size 100, 400 and 1000 and in each



case we computed the probability distribution of the number of cuts found by the BIC method and by the cross-validation method. The results are reported in Table 1. We also computed the total variation distance between true hazard and penalized hazard estimates in each case and reported the results in Table 2. We see that for  $n = 100$  both methods tend to be overpenalized as they find a majority of three breakpoints instead of four. As the sample size increases, the proportion of times the four breakpoints are found increases. Looking at the total-variation distance, we see that for both methods, the estimate becomes more and more accurate as the sample size increases. In general, the BIC criterion outperforms the cross-validation criterion both in terms of breakpoints detection and fitting of the hazard function.

One should note that the simulation scenario presented here makes it difficult to estimate the hazard function due to the low value of the hazard rates for  $t < 70$ . For a moderate sample size,  $n = 100$  for instance, very few observations will fall in each cut interval (only 8.5% in the interval  $(40, 50]$  for example) and therefore the method has difficulties to find all the cuts. The problem disappears as we increase the sample size. We considered other simulation settings where the proportion of observations falling into each cut interval was more balanced. This resulted in a very good performance of the estimator for small samples, both to detect the true number of cuts and to accurately fit the hazard function.

## 5.2 Simulations under a Weibull hazard model

We now consider the following Weibull model, where this time, the true hazard is a continuous function of time:  $\lambda(t) = a(t/b)^{a-1}/b$  where  $a = 5$  is the shape parameter and  $b = 60$  is the scale parameter. This gives an average time value of 55 and a time standard deviation of 12.6. The censoring distribution is also simulated as a Weibull variable but with shape parameter equal to 30 and a scale parameter equal to 60. This gives the same average percentage of observed failures (62%) as in the previous simulation setting.

As before we start with a single sample of size 100 generated from this model and we compute our adaptive ridge estimator using the same grid of cut points and the same grid of penalty values as in the previous scenario. The penalty value was chosen equal to 0.95 from the BIC criterion. Since we are estimating a continuous function of time it seems of interest to see how a smoother estimate would perform on this Weibull distribution. Our penalized likelihood can be easily modified to get a ridge estimate of the hazard by putting all the weights  $\mathbf{w}$  equal to 1 in Equation (3). This gives a simpler algorithm where the weights do not need to be updated and only a Newton-Raphson algorithm is performed on the parameter vector  $\mathbf{a}$ . However no simple criterion can be proposed to choose the penalty value in this setting and we arbitrarily chose a large value equal to 40 in order to force the estimator to be smooth. Plots of our adaptive ridge estimator, our ridge estimator and the true Weibull hazard are displayed in Figure 5. It is seen that two cuts are chosen for the adaptive ridge estimator which gives a fairly good fit of the true curve. However, as one would expect, the ridge estimator captures much more accurately the fluctuations of the curve. The resampling technique was used as before (100 samples) to compute the survival function along with its 95% confidence interval

in Figure 6. The time range for the figure was deliberately set to  $[0, 100]$  even though no times were observed beyond 60 due to censoring. The fit of the survival estimate is very accurate for the whole time range. After time 60 the piecewise constant modeling allows to interpolate the estimate which provides a good fit of the Weibull distribution with slightly larger confidence intervals. The Kaplan-Meier estimator is not shown on this figure because it gives similar result as for the piecewise constant hazard simulation scenario: a very similar fit to the curve and almost identical confidence intervals on the restricted time range  $[0, 60]$ . One should note however that our resampled estimator provides a much smoother fit than the stepwise shape of the Kaplan-Meier estimator and no interpolations can be provided after time 60 for the Kaplan-Meier estimator.

Finally, Monte-Carlo experiments were conducted to assess the quality of fit of our estimators for the Weibull hazard function. This was measured as before in terms of total variation distance between the true hazard and the adaptive ridge or the ridge estimator on the time interval  $[0, 60]$ . As an illustration, on the sample example of size 100 of Figure 5, the total variation distance equals 0.37 for the adaptive ridge estimate and 0.13 for the ridge estimate. It is important to note that a fixed penalty was used for every sample for the ridge estimator (equal to 40 as before) while the penalty was adaptively chosen from the BIC criterion as described in Section 3 for the adaptive ridge estimator. In terms of comparison this gives an initial advantage to the adaptive ridge estimator. Nevertheless the results reported in Table 3 show a clear advantage for the ridge estimator for every sample size. For  $n = 100$  the total variation error is 1.7 times bigger for the adaptive ridge estimator and for larger sample sizes it gets approximately 2 times bigger. These results indicate that if one aims at deriving smooth and accurate estimates of the hazard function, for prediction purposes for instance, one should favor the ridge version of our hazard estimator.

## 6 A real data analysis

We consider here the dataset from the Mayo Clinic trial in primary biliary cirrhosis (pbc) presented in Fleming & Harrington (1991). This dataset is available through the survival package of the R software. We focus our interest on time to death for the 424 patients of the dataset. The time variable was measured in days from inclusion until either death or liver transplantation or lost to follow-up. Only 38.5% of deaths are observed such that 61.5% of the observations are censored. The time variable ranges from 41 to 4795 days, so we decided to take as the set of all possible cuts the sequence of values going from 1 to 4800 by step of 10. As in the simulation study, the set of penalty terms was taken on the log scale, as the set of 100 equally spaced values ranging from  $\log(0.1)$  to  $\log(1000)$ . The penalty terms chosen from the BIC and cross-validation criteria are respectively equal to 1.23 and 1.63. This leads to one cut point for the BIC criterion and no cut point for the cross-validation criterion. Following the results from the simulation study, we decided to choose the former criterion. The corresponding estimate has one cut point at time 3050 such that the hazard estimate equals  $1.89 \cdot 10^{-4}$  for  $t \in (0, 3050]$  and equals  $3.84 \cdot 10^{-4}$  for  $t > 3050$ . The estimate and the regulation path for the penalty term are

displayed on Figure 7.

Finally, the bootstrap procedure is used to derive the survival estimate with its 95% pointwise confidence interval for the time to death. The curves are displayed on Figure 8 along with the Kaplan-Meier estimator and its 95% pointwise confidence interval. As in Section 5, the result from our estimator shows very little difference with the Kaplan-Meier estimator. With our bootstrap estimator, the median death time is estimated to approximately 3390 days and the 95% confidence interval for the survival at this time is approximately  $[0.43, 0.56]$ . The 25th percentile is estimated to approximately 1501 days and the 95% confidence interval for the survival at this time is approximately  $[0.70, 0.78]$ . With the Kaplan-Meier estimator the median is estimated at 3395 and the 95% confidence interval for the survival at this time is approximately  $[0.43, 0.57]$ , the 25th percentile is estimated to approximately 1462 days and the 95% confidence interval for the survival at this time is approximately  $[0.71, 0.79]$ .

## 7 Concluding remarks and extensions

In this article we proposed an innovative method to estimate the hazard rate in a piecewise constant model. The estimator is defined as the maximum of a penalized likelihood and allows to automatically detect the number and cuts location of the model and to estimate the hazard on each cut interval. A bootstrap procedure was also proposed in order to derive valid statistical inference taking both into account the variability of the estimate and the variability in the choice of the cut points. In order to select the penalty term we recommend using the BIC criterion as it seems to outperform the cross-validation criterion and it is also very fast to compute. Finally if one is interested in obtaining a smooth estimate of the hazard function, a small modification of the original estimator allows to derive a ridge version which has been shown to provide a very good fit to continuous survival distributions.

This work was established in the nonparametric setting of right censored data but many extensions can be considered. Including covariates in the model through a Poisson regression modeling for instance should be straightforward. As a matter of fact, since the method uses a penalized likelihood approach, no explicit estimators are available and even in the nonparametric setting the estimator is derived from the Newton-Raphson algorithm. In the nonparametric and regression settings, by modifying the likelihood formula, the method should also readily extend to truncation and to other types of censoring such as interval censoring. More difficultly it would be interesting to see how the penalized likelihood approach works in a frailty, joint modeling or cure model context. Using the L0 approach in these contexts amounts to fit a penalized parametric model which makes our method very appealing due to the nice properties of parametric models. Besides, our resampling method allows to derive smooth estimates of time dependent quantities of interest. As a result it is seen that our method nicely combines both the advantages of a parametric implementation and nonparametric fit of survival quantities.

The L0 approach was used to constrain two consecutive cuts in the piecewise constant hazard model to be equal. Interestingly, a different model could be proposed where

straight lines connect the consecutive cuts. In that model, the L0 approach could be derived by constraining two consecutive slopes of lines to be equal. In the same idea, spline hazard functions could also be constructed by penalizing further order derivatives of polynomial functions. All these extensions are left to future research.

## References

- AALLEN, O. O., BORGAN, Ø. & GJESSING, H. K. (2008). *Survival and Event History Analysis*. Statistics for Biology and Health. Springer.
- ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. & KEIDING, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. New York: Springer-Verlag.
- ANDERSEN, P. K., KLEIN, J. P., KNUDSEN, K. M. & TABANERA Y PALACIOS, R. (1997). Estimation of variance in cox’s regression model with shared gamma frailties. *Biometrics* **53**, 1475–84.
- ANTONIOU, A., PHAROA, P., SMITH, P. & EASTON, D. (2004). The boadicea model of genetic susceptibility to breast and ovarian cancer. *British Journal of Cancer* **91**, 1580–1590.
- CLAYTON, D., HILLS, M. & PICKLES, A. (1993). *Statistical models in epidemiology*, vol. 161. IEA.
- CLAYTON, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.
- COX, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* **34**, 187–220.
- FAREWELL, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* , 1041–1046.
- FLEMING, T. R. & HARRINGTON, D. P. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc.
- FROMMLET, F. & NUEL, G. (2016). An adaptive ridge procedure for l0 regularization. *PLoS ONE* **11**, 1–23.
- GROENEBOOM, P. (1996). Lectures on inverse problems. In *Lectures on probability theory and statistics*. Springer, pp. 67–164.
- GROENEBOOM, P. & WELLNER, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, vol. 19. Springer Science & Business Media.

- GRØN, R., GERDS, T. A. & ANDERSEN, P. K. (2016). Misspecified poisson regression models for large-scale registry data: inference for 'large n and small p'. *Stat Med* **35**, 1117–29.
- HOUGAARD, P. (1995). Frailty models for survival data. *Lifetime data analysis* **1**, 255–273.
- HVIID, A. & SVANSTRÖM, H. (2009). Antibiotic use and type 1 diabetes in childhood. *Am J Epidemiol* **169**, 1079–84.
- JENSEN, H. M., GRØN, R., LIDEGAARD, O., PEDERSEN, L. H., ANDERSEN, P. K. & KESSING, L. V. (2013). The effects of maternal depression and use of antidepressants during pregnancy on risk of a child small for gestational age. *Psychopharmacology (Berl)* **228**, 199–205.
- KESSING, L. V., THOMSEN, A. F., MOGENSEN, U. B. & ANDERSEN, P. K. (2010). Treatment with antipsychotics and the risk of diabetes in clinical practice. *Br J Psychiatry* **197**, 266–71.
- KLEIN, J. P., MOESCHBERGER, M., LI, Y., WANG, S. & FLOURNOY, N. (1992). Estimating random effects in the framingham heart study. In *Survival Analysis: State of the Art*. Springer, pp. 99–120.
- RIPATTI, S. & PALMGREN, J. (2002). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**, 1016–1022.
- RIZOPOULOS, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.
- RIZOPOULOS, D. et al. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**, 1–33.
- RONDEAU, V., FILLEUL, L. & JOLY, P. (2006). Nested frailty models using maximum penalized likelihood estimation. *Statistics in medicine* **25**, 4036–4052.
- RONDEAU, V., MATHOULIN-PELISSIER, S., JACQMIN-GADDA, H., BROUSTE, V. & SOUBEYRAN, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics* **8**, 708–721.
- RONDEAU, V., MAZROUI, Y. & GONZALEZ, J. R. (2012). frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software* **47**, 1–28.
- SIMON, N., FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2011). Regularization paths for cox’s proportional hazards model via coordinate descent. *J Stat Softw* **39**, 1–13.
- SUN, J. (2007). *The statistical analysis of interval-censored failure time data*. Springer Science & Business Media.

SY, J. P. & TAYLOR, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics* **56**, 227–236.

THERNEAU, T. M. & GRAMBSCH, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

TSIATIS, A. A. & DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* , 809–834.

VAN HOUWELINGEN, H. C., BRUINSMA, T., HART, A. A. M., VAN’T VEER, L. J. & WESSELS, L. F. A. (2006). Cross-validated cox regression on microarray gene expression data. *Stat Med* **25**, 3201–16.

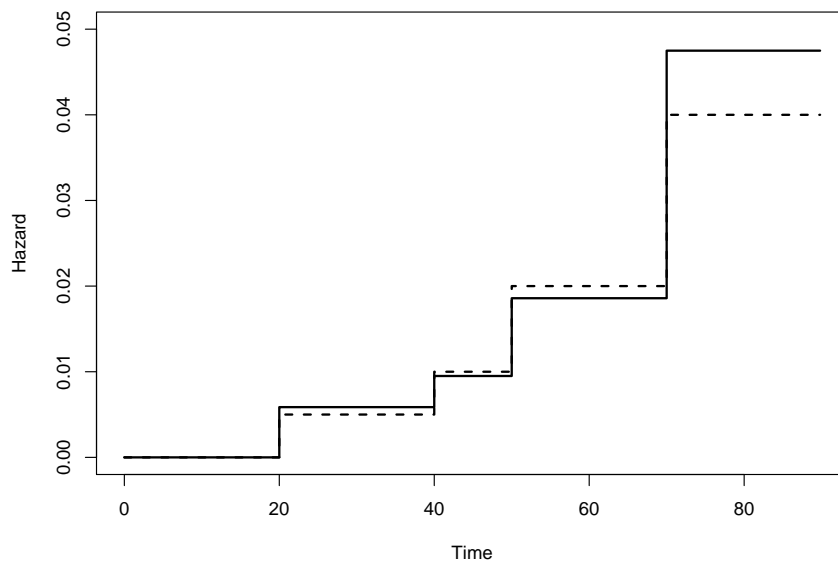


Figure 1: True hazard rate function (dashed line) and unpenalized hazard rate estimate computed at the true cuts (solid line).

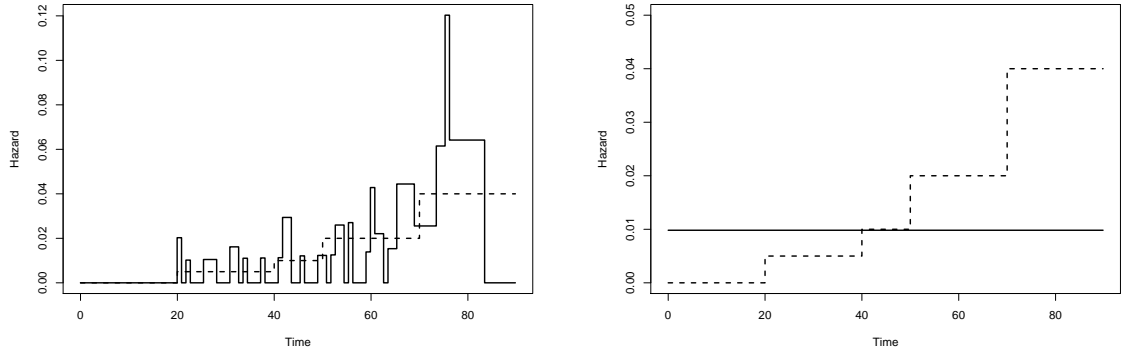


Figure 2: Penalized hazard rate estimates computed using a penalty equal to 0.1 (left panel) and a penalty equal to 1000 (right panel). Dashed line: true hazard rate. Solid lines: penalized hazard rate estimates.

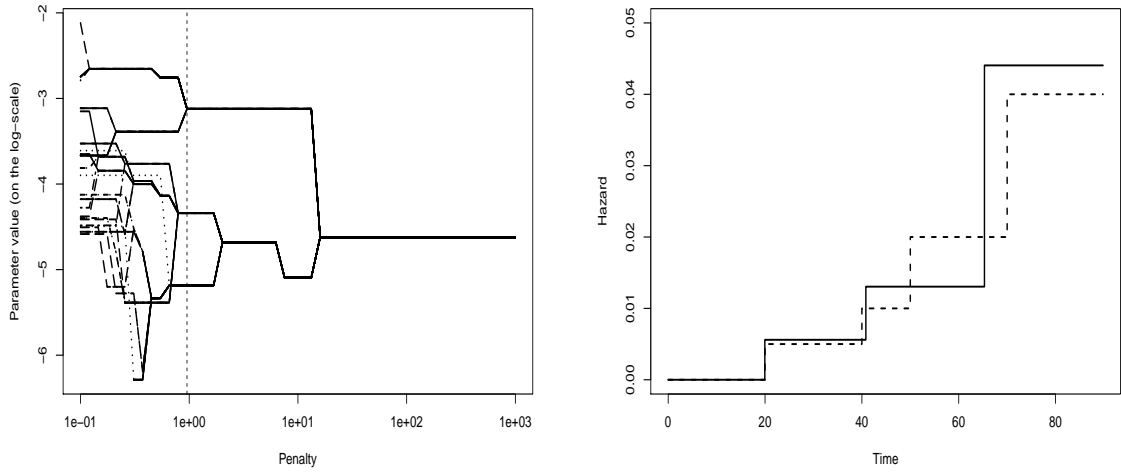


Figure 3: Regularization for the choice of the penalty term using either the BIC or cross-validation criterion (left panel). Dashed line: penalty term obtained from both criteria. Penalized hazard rate estimate (right panel).

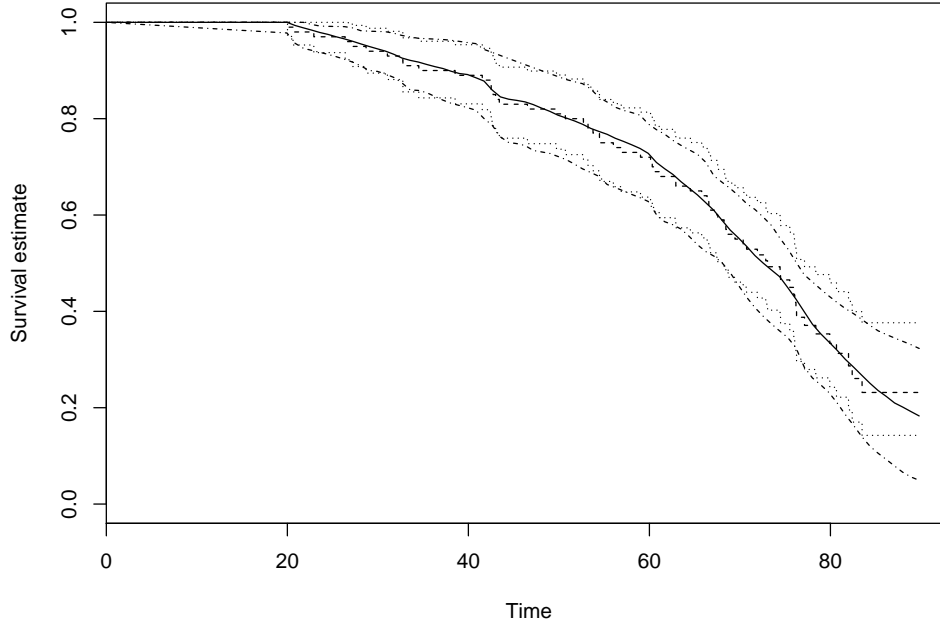


Figure 4: Estimates of the survival function for the piecewise constant hazard scenario. Dashed line: Kaplan Meier estimator along with its 95% pointwise confidence interval (dotted lines). Solid line: bootstrapped adaptive ridge estimator along with its 95% pointwise confidence interval (dot dash lines).

Table 1: Proportions of the number of cuts found by the BIC (left panel) and cross-validation (right panel) criteria for different sample sizes.

Number of cuts	Proportions found for:			Number of cuts	Proportions found for:		
	$n = 100$	$n = 400$	$n = 1\,000$		$n = 100$	$n = 400$	$n = 1\,000$
0	0.000	0.000	0.000	0	0.000	0.000	0.000
1	0.000	0.000	0.000	1	0.075	0.000	0.000
2	0.202	0.005	0.000	2	0.338	0.032	0.000
3	0.363	0.328	0.038	3	0.323	0.280	0.045
4	0.202	0.375	0.737	4	0.105	0.352	0.615
5+	0.233	0.292	0.225	5+	0.158	0.337	0.340



Table 2: Mean total variation distance between true hazard and penalized estimated hazard obtained from the BIC and cross-validation (CV) criteria for different sample sizes in the piecewise constant hazard scenario.

	$n = 100$	$n = 400$	$n = 1\,000$
BIC	0.362	0.176	0.085
CV	0.370	0.184	0.092

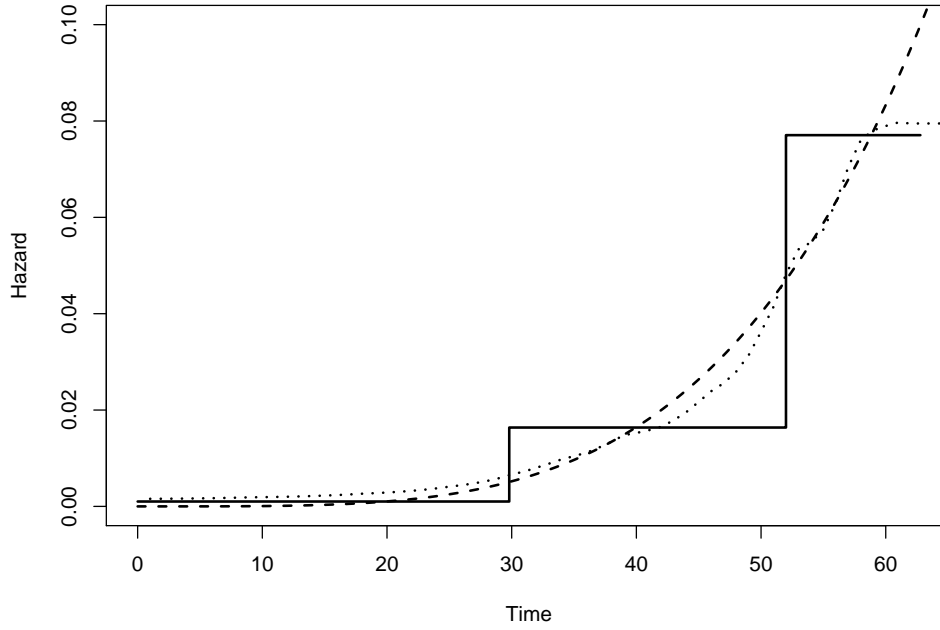


Figure 5: Penalized hazard rate estimates for the Weibull scenario. Dashed line: true hazard. Solid line: adaptive ridge estimator. Dotted lines: ridge estimator.

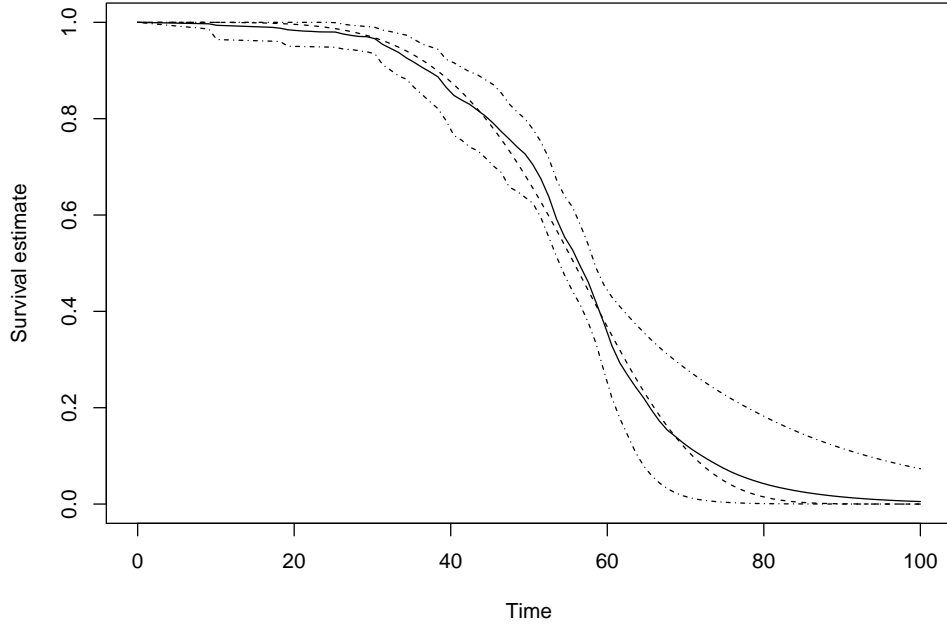


Figure 6: Estimates of the survival function for the Weibull scenario. Dashed line: true hazard. Solid line: bootstrapped adaptive ridge estimator along with its 95% pointwise confidence interval (dot dash lines).

Table 3: Mean total variation distance between true hazard and penalized estimated hazard obtained from the adaptive ridge estimator and the ridge estimator for different sample sizes in the Weibull scenario.

	$n = 100$	$n = 400$	$n = 1\,000$
Adaptive Ridge	0.347	0.228	0.172
Ridge	0.204	0.115	0.086

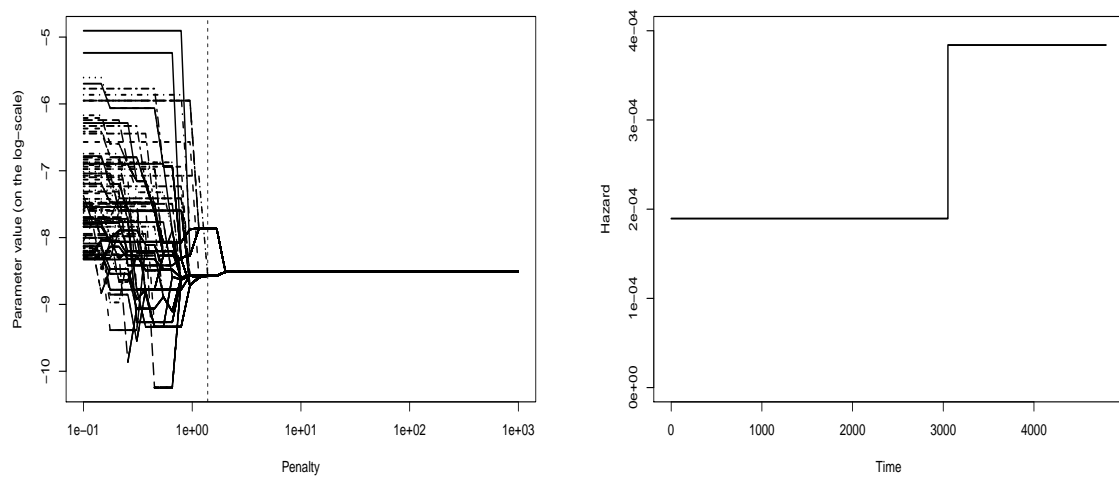


Figure 7: Regularization for the choice of the penalty term using the BIC criterion on the pbc data (left panel). Dashed line: penalty term obtained from this criterion. Penalized hazard rate estimate of death on the pbc data (right panel).

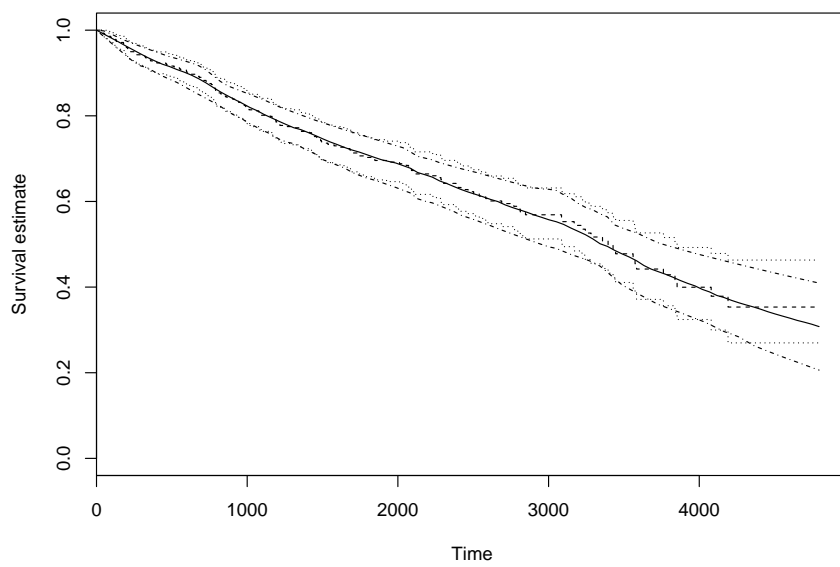


Figure 8: Estimates of the survival function on the pbc data. Dashed line: Kaplan Meier estimator along with its 95% pointwise confidence interval (dotted lines). Solid line: bootstrapped adaptive ridge estimator along with its 95% pointwise confidence interval (dot dash lines).